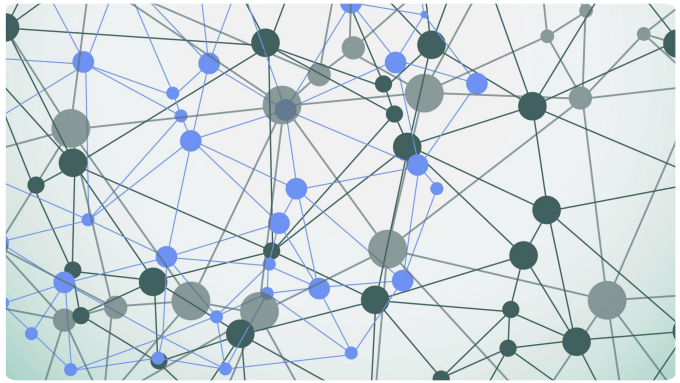
UCI nlp

Revisiting Evaluation of Knowledge Base Completion Models

Pouya Pezeshkpour, Yifan Tian, Sameer Singh



Revisiting Evaluation of KB Completion

Evaluating KG Completion

Shortcomings

1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC

Revisiting Evaluation of KB Completion

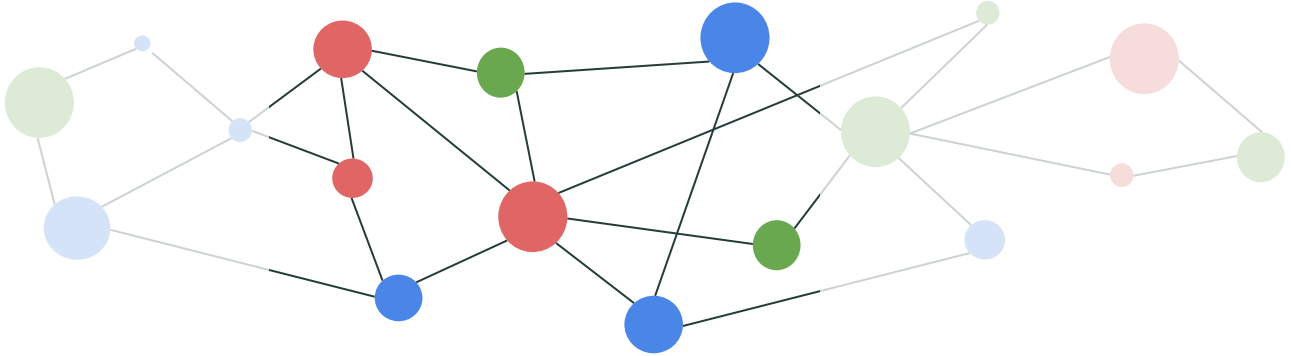
Evaluating KG Completion

Shortcomings

1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC

Knowledge Graphs



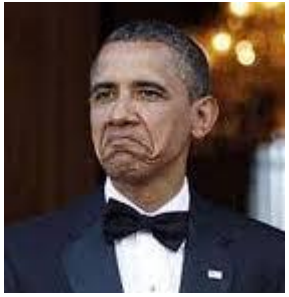
Unstructured text

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ (listen);[1] born August 4, 1961) is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, Barack Obama was the first African-American president of the United States. He previously served as a U.S. senator from Illinois from 2005 to 2008 and an Illinois state senator from 1997 to 2004.

Structured text

44th President of the United States	
In office	January 20, 2009 – January 20, 2017
Vice President	Joe Biden
Preceded by	George W. Bush
Succeeded by	Donald Trump
United States Senator from Illinois	
In office	January 3, 2005 – November 16, 2008
Preceded by	Peter Fitzgerald
Succeeded by	Roland Burris

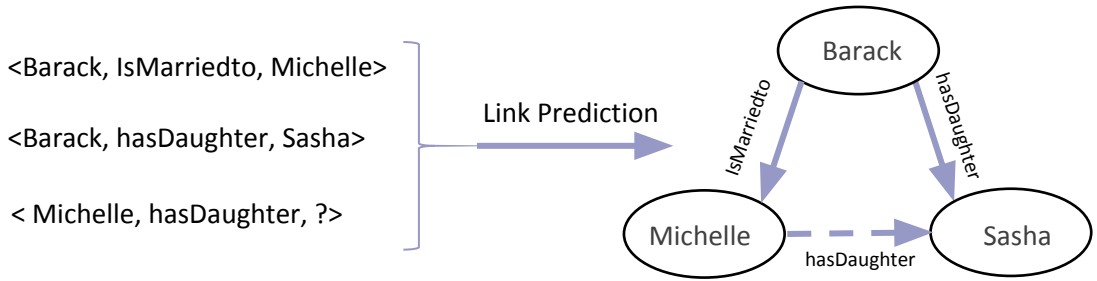
Images



Shortcomings of Knowledge Graphs

Because of the way they are created:

- There are many missing facts
  - ➡ KG Completion



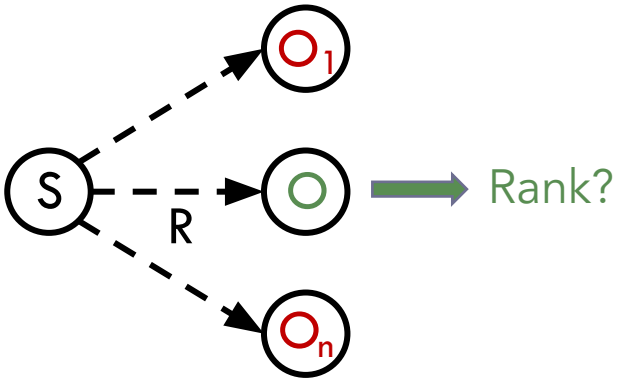
- The factuality of non-existent links is unknown (open-world assumption)
  - ➡ Use of ranking for evaluation

But, real-world application mostly care about an information being True or False and not the ranking

Knowledge Graph Completion Evaluation

- Two evaluation approach for scoring target triple <S, R, O>:

1. Ranking Metrics:



Doesn't correspond to real-world use case in many instances

2. Triple Classification:

Learning thresholds τ_R by randomly choosing negative samples for validation data

Revisiting Evaluation of KB Completion

Overview of Knowledge Graph Completion

Shortcomings

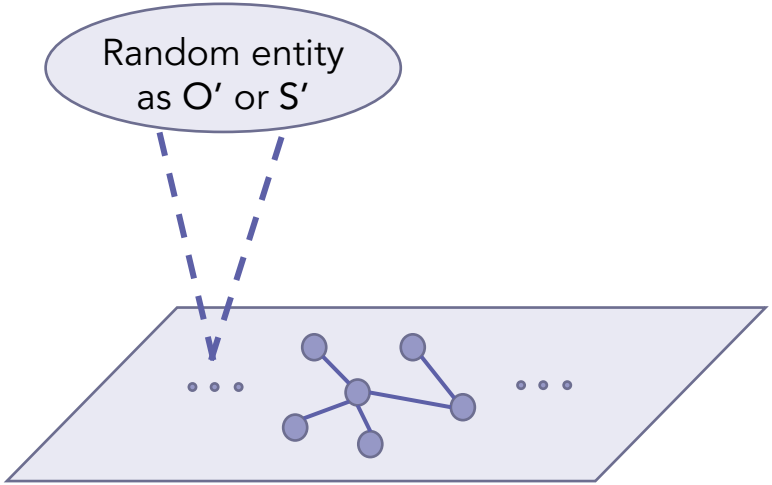
1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC

Negative Sampling

- We consider 3 different negative sampling for target triple <S,R,O>:

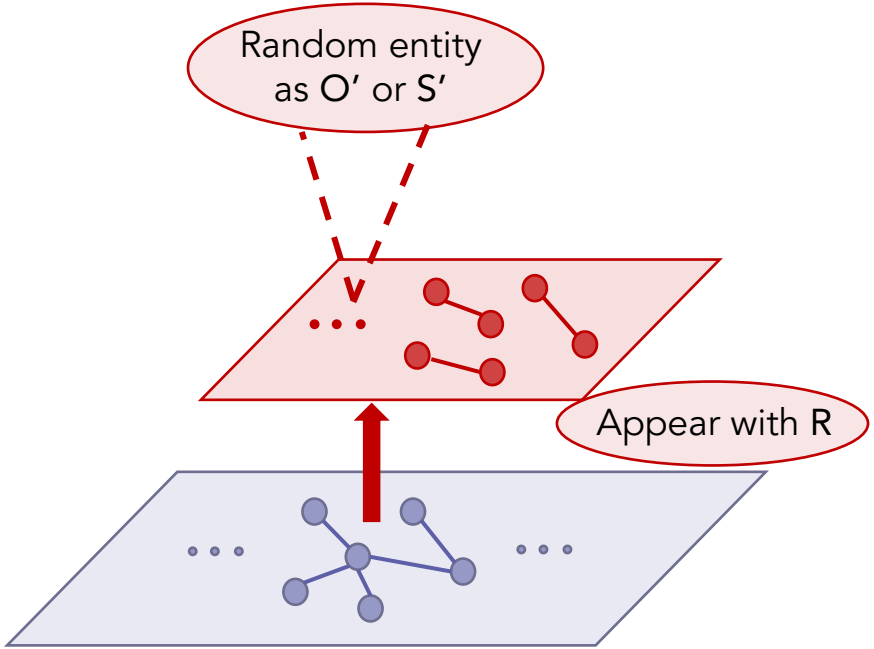
1. Random Sampling:



Negative Sampling

- We consider 3 different negative sampling for target triple <S,R,O>:

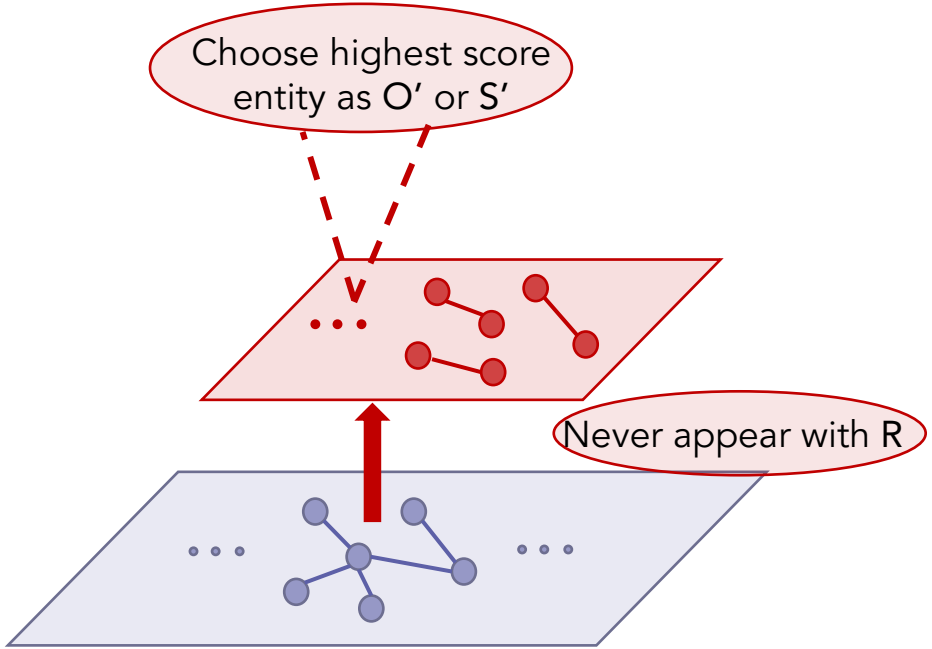
2. Constraint Sampling:



Negative Sampling

- We consider 3 different negative sampling for target triple <S,R,O>:

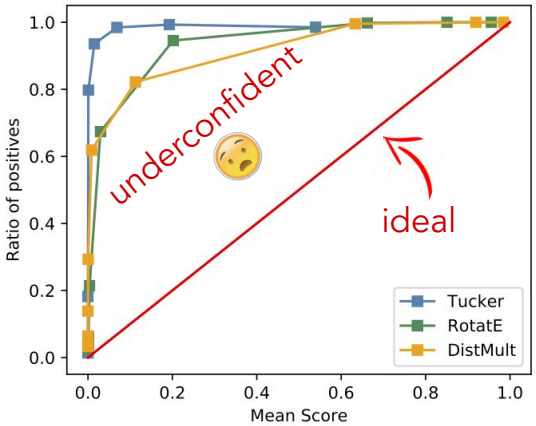
3. Careful Sampling:



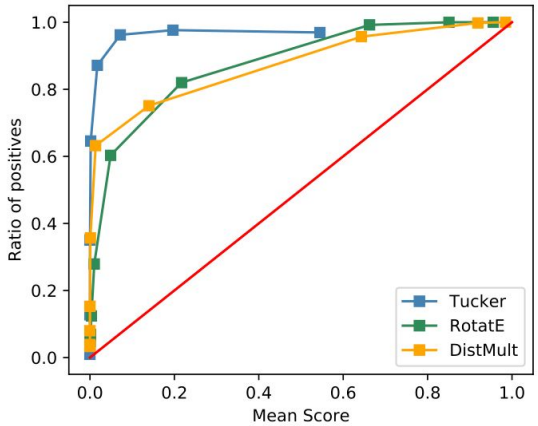
Calibration

- Calibration study is not well defined

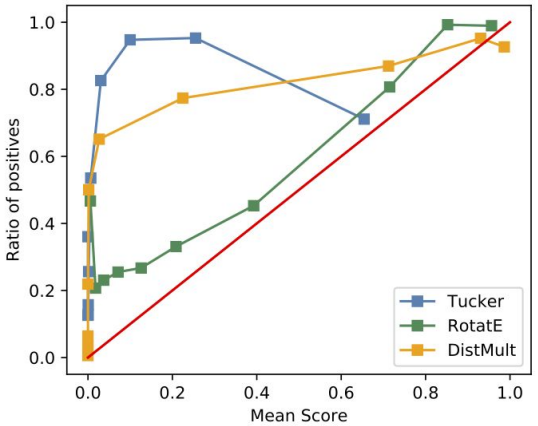
1. Random sampling



2. Constrained sampling



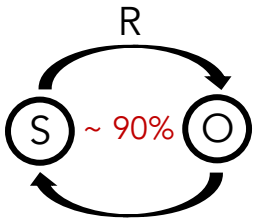
3. Careful sampling



Extremely different conclusions
for different negative samplings

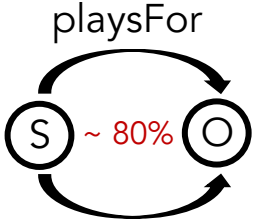
Semi-Inverse Relations

WN18RR

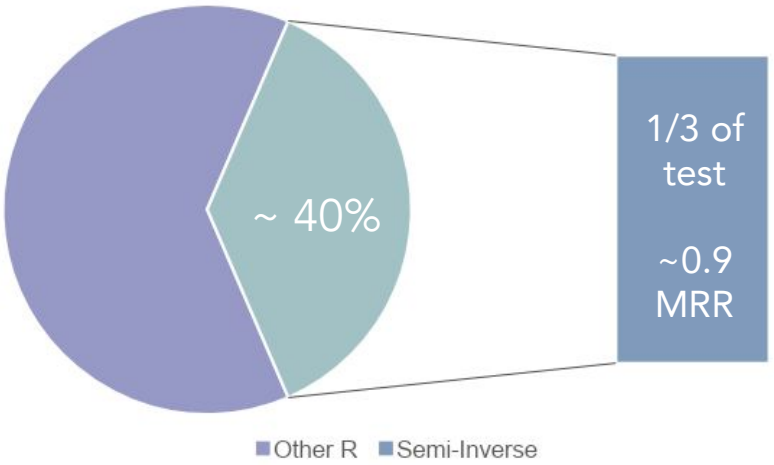


Ranking metrics do not reflect reasoning power

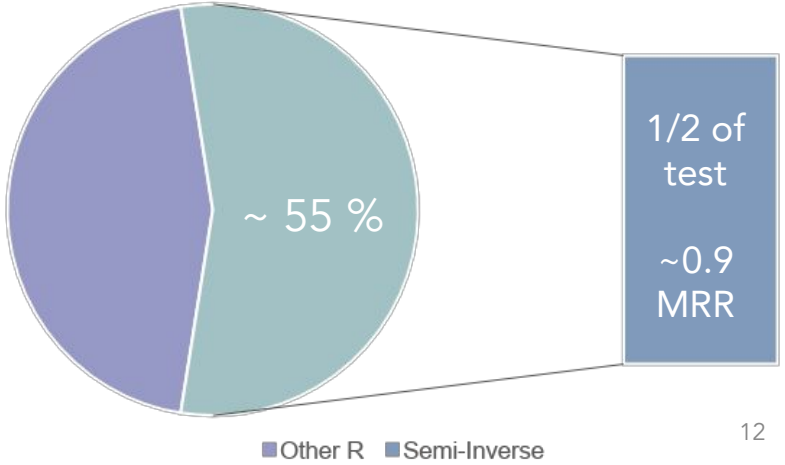
YAGO3-10



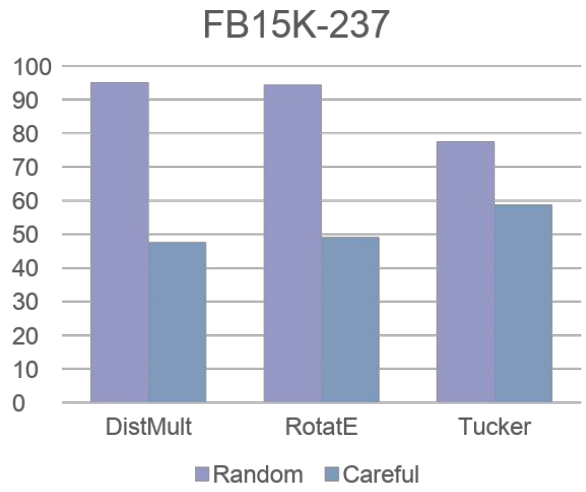
Train data

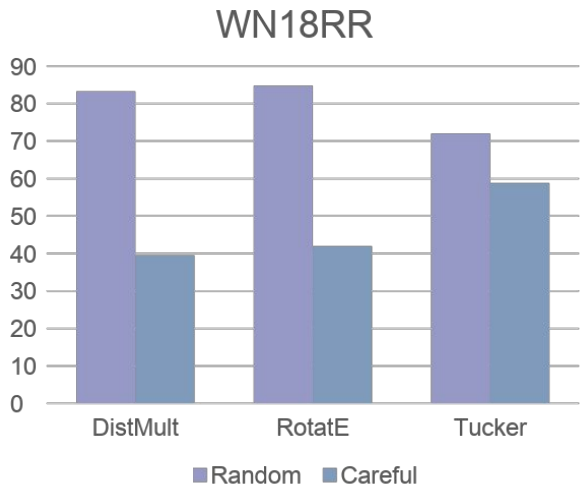


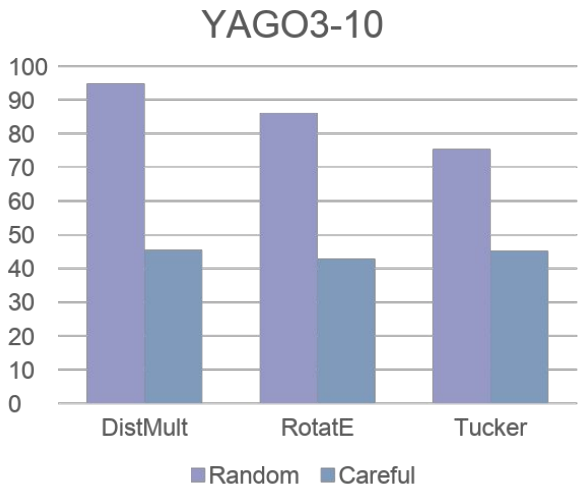
Train data



Triple Classification Robustness







Careful negative sampling results in a dramatic drop

Results are around 90 %

Revisiting Evaluation of KB Completion

Overview of Knowledge Graph Completion

Shortcomings of evaluation metrics

1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC

YAGO3-TC Dataset

What is Our goal?

- Create a benchmark that align with real-world application
- Properly differentiate between models
- Capture reasoning powers

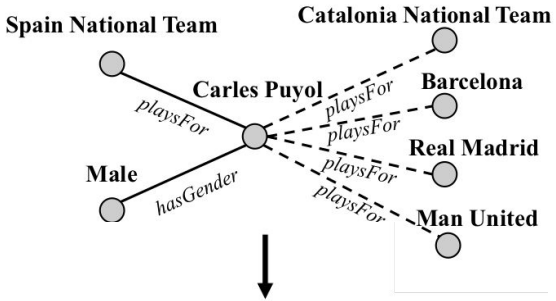
What are existing challenges?

- The knowledge graphs are not complete
- There so many non existent links
- Identifying the factuality of missing information is hard

YAGO3-TC Creation

1. Randomly choose a subset of YAGO3-10 test
2. Identify top scoring triples from the models that are unknown to be true
3. Filter triples
4. Crowdsourcing pipeline

Crowdsourcing Pipeline

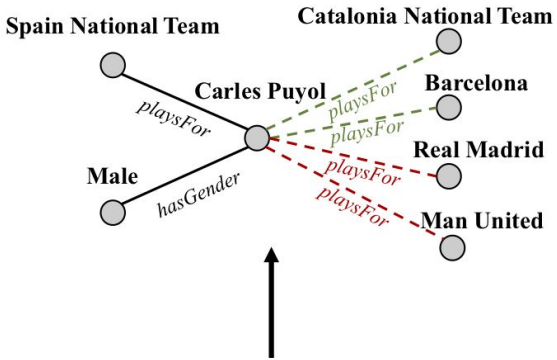


Initial Annotations (Round 1)

Carles Puyol plays for?

☐ Catalonia National Team	✓	✓	×
☐ Barcelona	✓	✓	✓
☐ Real Madrid	✓	×	×
☐ Man United	×	×	×

✓ Catalonia National Team
✓ Barcelona
× Man United
? Real Madrid



Selective Reannotation (Round 2)

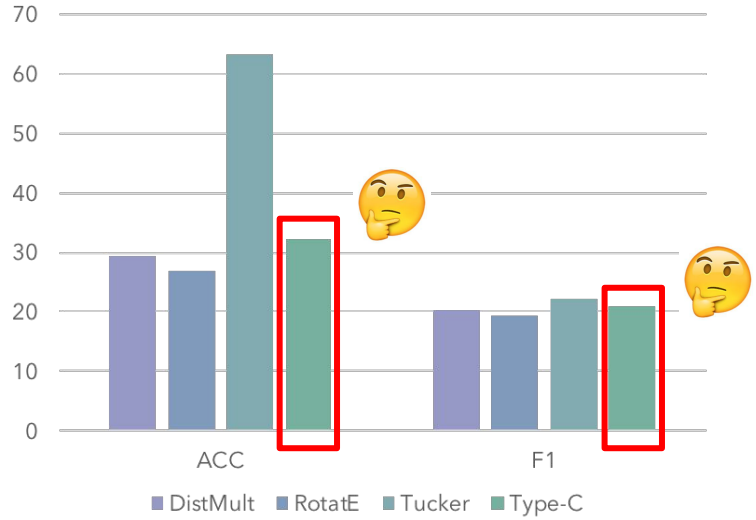
Carles Puyol plays for?

☐ Real Madrid	×	×

- ~ 30 K triples
- ~ 10% positives

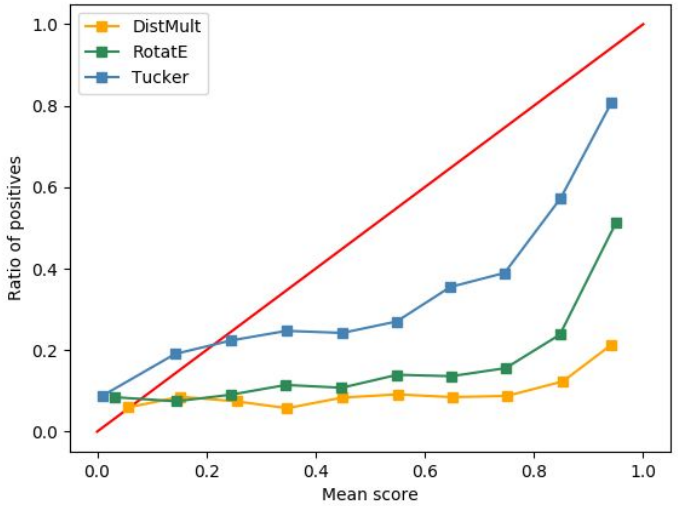
Evaluation Using YAGO3-TC

- Triple classification



- SOTA models perform poorly
- Huge drop in accuracy

- Calibration



- Reverse order of models
- Overconfident

Discussion

- Ranking metrics are not very trustworthy
- Triple classification is not robust
- Real-world adoption of KG needs better evaluation techniques
- YAGO3-TC is the first step toward this goal

We propose a web-hosted evaluation platform to update YAGO3-TC using new KG completion models

Revisiting
Evaluatio
KB Comp

mpletion

nverse relations
ration
classification
ss

Introducing YAGO3-TC

Thank You!

Website (code, data and leaderboard):

pouyapez.github.io/yago3-tc/

Contact me: **pezeshkp@uci.edu**